BioE/MCB/PMB C146/246, Spring 2003

Problem Set 3: General Gap Penalties, Advanced Dynamic Programming, Substitution Matrices

Due 10 Feb 03, 5:00 pm PST by email to derek@rana.lbl.gov

1. 15 points

Sequence Alignment with Affine Gaps

Revise your alignment program from the last problem set using the following parameters:

Identity	+4
Transition	–2
Transversion	–4
Gap	–8 (first position), –1 (each gap position)

Given the sequences

ATTTTAAGCGCATACCGC
TCGCAAATATAC

Perform a global alignment on the two sequences and report their score. Attach all dynamic programming matrices used (without tracebacks) to your email as `sid_ps3_1.txt`

2. 20 points

Perform any two of the following alignments, using the scoring matrix from Problem Set 2. You may use a program to assist you, though implementations are not required. Use gap penalties of –5 at each position, except where otherwise specified.

(i)	Repeated matches, threshold 10	HECYDWH	and	HEWGH
(ii)	Hirschberg/Myers/Miller alignment	HECYDWH	and	HEWGH
(iii)	Sub-optimal global alignment	HEAGAWGHE	and	PAWHEA
(iv)	Global alignment with three-parameter gaps (–8, –1, –2) for the sequences			

SSFTLT and SCHKDIL

Include dynamic programming matrices, traceback paths, alignments and scores in your answer. Attach the dynamic programming matrix (without tracebacks) to your email as `sid_ps3_2.txt`

3. 5 points

Why are sub-additive gap penalties used? Give (at least) two reasons.

Sub-additive gap penalties lower the time required for computing gap penalties. In the special case of affine gaps, they reduce the gap computation to constant time, instead of O(*n*) time.

Sub-additive gap penalties provide a better biological model for gaps than per position gap penalties. Gaps are difficult to open in protein sequences, because of constraints on protein structure. However, once a gap is allowed in a structurally flexible region, it should be much easier to extend the gap. Therefore, it makes sense to penalize the initiation of a gap more heavily.

4. 5 points

Compare and contrast the construction and features of the BLOSUM and PAM series of matrices. Mention the strengths and weaknesses of each.

The construction of a PAM matrix begins with a group of aligned sequences that are 99% similar. A phylogenetic tree for the sequences is inferred. PAM assumes a Markovian mutational process, where each subsequent mutation of a residue is not affected by any previous mutations. From the alignments, the relative mutability of each residue (normalized to 1%) is calculated by dividing the number of accepted changes by the exposure to mutation. The number of accepted changes can be calculated by the number of mutations in the extant sequence compared to the inferred ancestor. Mutability between two residues a and b is calculated as follows: $M_{ab} = \text{freq}(ab) \times \text{mutability}(a) / \text{freq}(a)$. PAM^n matrices are generated by matrix multiplication of the M matrices, n times. The higher the number if the PAM matrices, the longer the evolutionary time and the less similar the sequences.

There are several problems with PAM matrices. First, they assume Markovian mutational processes. This may not be accurate because certain residues may be hyper-mutable (*eg.* regions of protein that are unstructured) and others hypo-mutable (*eg.* the catalytic core), which would not be accounted for in the PAM matrix. Second, it assumes that residues are independent, which isn't necessarily true for proteins. This information can't be incorporated into any scoring matrix, though, without detailed information about the protein's structure, which isn't feasible. Third, all calculations are based upon the original phylogeny, which is an approximation. Fifth, the mutations seen in a PAM matrix are biased toward those seen in short evolutionary time since the starting sequences were so similar.

BLOSUM matrices are constructed from groups of ungapped aligned sequences whose percent identity meets a particular threshold similarity. From the alignment, the joint probability of any pair of residues in a column is computed. Dividing by the product of the marginal probabilities for the individual residues, a likelihood ratio can be computed for observing two residues aligned due to common ancestry versus by chance.

BLOSUM matrices have too important advantages over PAM matrices: they work better in practice and they are derived from actual substitutions rather than inferred ancestral sequences. BLOSUM doesn't require any evolutionary model for its construction, which can be a disadvantage since we are attempting to infer evolutionary relationships between sequences when we implement the matrices.

5. 5 points
A 1-PAM matrix changes on average of 1% of amino acids. Does a 2-PAM matrix change on average 2%? Explain.

Since multiple mutations may occur in the same position, slightly less than 2% of the amino acids will change.

6. 5 points
For alignments performed with PAM matrices, explain the meaning of a substitution score and the score of the alignment.

The substitution score of a PAM matrix represents a likelihood ratio: the likelihood that two residues have common ancestry over the period of time represented by the PAM matrix, versus the likelihood that the residues are aligned by chance.

The score for the alignment represents the overall likehood ratio for the entire alignment to be descended from a common ancestor over the given PAM distance, divided by an alignment formed by chance. As the sum of the substitution scores for each pair of residues in the alignment, it represents joint likelihoods, assuming that each position in the alignment is independent.

7. 5 points
Why are gap parameters NOT estimated the same way as substitution matrix parameters?

Gap parameters are not estimated the same way as substitution score for several reasons. Generalized gap penalties use varying costs for different regions within a gap. Therefore, the assumption of positional independence used to construct substitution matrices does not hold for gaps. In addition, note that a gap doesn't actually exist outside of an alignment; instead, gaps are indicators of insertions or deletions. Calculating the frequency of an alignment between a residue x and a gap is actually measuring the frequency with which x is inserted or deleted from a sequence (these two possibilities are indistinguishable).

8. 15 points
Given the following BLOCK (multiple sequence alignment of proteins):

```
MMKE
MKKE
IKIE
MEME
IMKI
IKKE
MKME
IKKE
MKME
IKKE
```

(A) (10 points) Compute the joint probabilities q_{ij} and the marginal probabilities p_i for each i, j in the amino acid alphabet.

Frequency table

M	5	2	3	0
I	5	0	1	1
K	0	7	6	0
E	0	1	0	9

Marginal probabilities: **M = 0.25** **I = 0.175** **K = 0.325** **E = 0.25**

Joint probabilities

	M	I	K	E
M	0.078			
I	0.156	0.056		
K	0.178	0.033	0.2	
E	0.011	0.05	0.039	0.2

Product of marginal probabilities

	M	I	K	E
M	0.063			
I	0.088	0.031		
K	0.163	0.114	0.106	
E	0.125	0.088	0.163	0.063

(B) (5 points) Compute the BLOSUM matrix for this BLOCK.

Log_2-Odds Matrix

	M	I	K	E
M	0.316			
I	0.83	0.859		
K	0.13	-1.77	0.921	
E	-3.49	-0.81	-2.06	1.678

9. 15 points

Given the initial mutability matrix below, calculate the corresponding 3-PAM matrix. Normalize your answer such that each row and column sums to 1000.

	S	T	V
S	990	7	3
T	7	993	0
V	3	0	997

3-PAM

	S	T	V
S	970.47	20.64	8.88
T	20.64	979.29	0.06
V	8.88	0.06	991.05